

Ordering as privileged information

Tom Vacek

University of Minnesota, Minneapolis, MN

vacek@cs.umn.edu

July 1, 2016

Abstract

We propose to accelerate the rate of convergence of the pattern recognition task by directly minimizing the variance diameters of certain hypothesis spaces, which are critical quantities in fast-convergence results. We show that the variance diameters can be controlled by dividing hypothesis spaces into metric balls based on a new order metric. This order metric can be minimized as an ordinal regression problem, leading to a LUPI application where we take the privileged information as some desired ordering, and construct a faster-converging hypothesis space by empirically restricting some larger hypothesis space according to that ordering. We give a risk analysis of the approach. We discuss the difficulties with model selection and give an innovative technique for selecting multiple model parameters. Finally, we provide some data experiments.

1 Introduction

Learning using Privileged Information (first proposed by Vapnik et. al. [1]) seeks to bring in privileged information to assist the learner. This information is privileged because the learner may use it choose a hypothesis, but the privileged information will be unavailable making decisions based on the hypothesis.

This paper proposes a LUPI method that directly minimizes the variance diameters of the hypothesis spaces under consideration, an essential quantity in fast-convergence literature [2, 3]. This approach applies to discriminant-based hypotheses spaces where predictions are derived from thresholding the the discriminant value, which should be totally orderable. We show that the discriminant ordering defines equivalence classes for the elements of the space, and these classes are directly related the the variance diameters we seek to control. If we could restrict the hypothesis space to just the good equivalence classes, we would reduce the variance diameters and improve the speed of convergence.

Selecting a good equivalence class requires some external definition of desirable order. This is privileged information. This raises a natural question: *What is a desirable order?* If ordering information were provided by an oracle according to some true distribution, then *any* ordering would provide desirable variance diameters. However, since an empirical ordering is the best we can hope for, a good ordering is one which has favorable convergence properties in ordinal regression.¹ Low ordinal loss is sufficient, while more general characterizations may be possible.

From another perspective, orderings according to conditional probability $P(Y|X)$ have great appeal, as models that provide good estimates of conditional probability allow broader application than ones that do

¹We would want the ordering to correspond to a good hypothesis for the pattern recognition problem, though this is irrelevant to the *rate* of convergence.

not. Moreover, confidence seems to be a sliver of common ground between human and machine learning. Thus, we consider orderings which seem to bear some relationship to conditional probability, though quite loose. While a total ordering is best for controlling variance diameters, two independent orderings for each class can be shown to be very nearly as good. This arrangement allows us to expand the scope of tasks where useful privileged information is available.

2 What is order?

We can't hope to provide any kind of overview of the field of order statistics. Web search made this a lucrative and popular field. Nevertheless, we believe our starting point is novel:

Definition 1. *Order* is any property of a set of real numbers that is invariant under any invertible increasing transformation.

Ordering naturally defines equivalence classes on hypotheses spaces. Suppose some distribution P generates feature vectors $X \in \mathbb{R}^d$, and suppose h_1 and h_2 are hypotheses in a space $\mathcal{H} : \mathbb{R}^d \rightarrow \mathbb{R}$. If an increasing function m exists so that for all $X \sim P$, $m(h_1(X)) = h_2(X)$,² then h_1 and h_2 are in the same equivalence class.

The thread which connects order to the pattern recognition task is the growth function, which measures the number of possible labelings of a set of points of size n by a 0/1-hypothesis set $\mathcal{H}^{0/1}$. We assume that $\mathcal{H}^{0/1}$ is defined by characteristic functions of real-valued functions, as in $\mathcal{H}^{0/1} = \{\xi(h(X) - t) : h \in \mathcal{H} : \mathbb{R}^d \rightarrow \mathbb{R}, t \in \mathbb{R}\}$. We observe that if \mathcal{H} contains a restricted number of order equivalence classes, then the growth function of \mathcal{H} is also restricted. We propose that the relationship can be made precise by the machinery of variance-based risk bounds.

As a thought experiment, consider the variance diameter of h_1 and h_2 when combined with an appropriate 0/1 loss function, and restricted to one class. More formally, suppose now that P generates feature vectors and labels $Y \in \pm 1$ jointly.

Definition 2. *Zero-one loss* is:

$$l^{01}(\hat{y}, y) = \begin{cases} 1 & \hat{y} > 0, y \leq 0 \\ 1 & \hat{y} \leq 0, y > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

3

Then for any h_1 and h_2 in the same order equivalence class for for any t_1 and t_2 , it is easy to show that

$$\mathbb{E}_{P|Y=1} |l^{01}(h_1(X) - t_1, Y) - l^{01}(h_2(X) - t_2, Y)| \quad (2)$$

$$\leq \mathbb{E}_{P|Y=1} |l^{01}(h_1(X) - t_1, Y) - l^{01}(h_2(X) - t_2, Y)| \quad (3)$$

²We might relax this to holding on a set of full measure.

³ As a matter of boilerplate, we use loss functions as are commonly defined in machine learning literature: a loss function takes two arguments: a prediction and a label; however, we may omit those where obvious to avoid clutter. Loss functions are uniquely identified by a superscript, unless intended to be taken generally. The expectation of the loss $\mathbb{E}_{X,Y \sim P}[l(h(X), Y)]$ is depicted as $L(h)$, and the empirical risk on a finite set of size n is depicted as $L_n(h)$. There are numerous more measurability and existence assumptions that we will not cover here.

Except for the restriction to a single class, this is just the relationship for variance diameters⁴ that is needed for fast rates.

There are two tasks required to extend the thought experiment to a real learning formulation. First, the relationship needs to apply in both classes simultaneously. The complicating factor to just adding the two per-class relationships is that the absolute-value arguments in the two respective RHS's could have opposite signs and cancel out. This can happen when the decision threshold falls in very different places relative to the ordering. We fix this by requiring that the loss in the two classes be balanced as a constraint on valid solutions. In effect, this is a constraint on the location of the decision boundary in the equivalence class definition, preventing the situations where there is cancellation.

More significantly, there is no access to the equivalence classes based only on empirical information. This is the majority of the analysis in the remainder of the paper. In short, we relax the notion of the equivalence class to metric balls, and then we bound the deviation of empirical balls from their true diameter.

2.1 Ordering metric

We define a metric on orderings so that two hypotheses are in the same equivalence class if their metric distance is 0. The metric, when shown to have favorable properties, allows us to create balls of restricted variance diameter based on a finite sample using ordinary empirical risk minimization.

Definition 3. Let \mathcal{M} be the set of all increasing continuous functions.

Definition 4. Let P generate vectors $X \in \mathbb{R}^d$, and consider any two functions $h_1, h_2 : \mathbb{R}^d \rightarrow \mathbb{R}$. Then the order distance between h_1 and h_2 is

$$D(h_1, h_2) = \sup_{t \in \mathbb{R}} \inf_{m \in \mathcal{M}} \mathbb{E}_P[l^{01}(m \circ h_1(X) - t, h_2(X) - t)]$$

The metric axioms (up to equivalence class elements) are not hard to check; the invertability of m is indispensable here.

While many definitions would satisfy the metric axioms, we chose this definition for two reasons. First, it is a direct extension of the result we saw for equivalence classes. If $D(h_1(X), h_2(X)) \leq d$, then (3) holds with d added to the right-hand side. Second, the fact that 0/1 loss is an underlying component allows us to borrow a great deal from the standard results in machine learning.

Since we have no access to true probabilities in a statistical learning setting, to proceed we have to derive a way to get access to D . We accomplish this by extending D to measure the order distance of a function h to some *ground truth* ordering instead of another function. The key observation is that functions within D_0 of the ground truth ordering are within $2D_0$ of each other by the triangle inequality—an empirical version of the equivalence classes we set out to find. We will redefine this extension of D as L^{iso} for clarity:

⁴ Fast converging bounds require the following: Denote by $h' \in \mathcal{H}$ the minimum of the true risk over \mathcal{H} . The following must hold uniformly for all $h \in \mathcal{H}$:

$$\begin{aligned} & \text{Var}[l^{01}(h(X), Y) - l^{01}(h'(X), Y)] \\ & \leq \mathbb{E}[l^{01}(h(X), Y) - l^{01}(h'(X), Y)] \end{aligned} \tag{4}$$

Our presentation is slightly different because the variance can be upper bounded by the expectation of the absolute value and we have not specialized to h' . Note that many presentations of fast convergence can lead an in-attentive reader to believe that h' must be the Bayes rule. This is true if one is using the Mammen-Tsybakov noise conditions to establish the desired variance relationship, but not necessary if the relationship can be established another way, as we do here.

Definition 5. Suppose P jointly generates feature vectors X and real-valued labels Y . Let $\mathcal{H} : \mathbb{R}^d \rightarrow \mathbb{R}$ be some hypothesis space.

$$L^{iso}(h) = \sup_{t \in \mathbb{R}} \inf_{m \in \mathcal{M}} \mathbb{E}_P[l^{01}(m \circ h_1(X) - t, Y - t)]$$

The task at hand is to analyze risk bounds for L^{iso} that allows us, with high probability, to identify functions where $L^{iso}(h) \leq D_0$. based on a finite sample. The key observation is that the infimum ($m \in \mathcal{M}$) and supremum ($t \in \mathbb{R}$) in the definition of L^{iso} can be handled by a uniform bound on the deviations of a loss function over the joint space $\mathcal{H} \times \mathcal{M} \times \mathbb{R}$. We define that loss function to be l^{thr} :

Definition 6.

$$l^{thr}(\hat{y}, y, t) = l^{01}(\hat{y} - t, y - t)$$

We will show that if \mathcal{H} has bounded level VC-dimension,⁵ then the loss class for l^{thr} over $\mathcal{H} \times \mathcal{M} \times \mathbb{R}$ also has bounded VC dimension, so we can derive uniform risk bounds for empirical risk minimization.

We observe that level VC-dimension already satisfies an order invariance. That is, the inclusion of \mathcal{M} doesn't have any affect on the estimated growth function.

Lemma 1. Let $\mathcal{MH} = \{m \circ h : h \in \mathcal{H}, m \in \mathcal{M}\}$. Then the level VC dimension of \mathcal{MH} is the same as for \mathcal{H} .

The proof is a simple shattering argument. Obviously, $\mathcal{MH} \supseteq \mathcal{H}$, so the VC-dimension is not less. It is not more because any rule $m \circ h - t > 0$ can be replicated by $h - t' > 0$ for an appropriately chosen t' .

These two observations give the following theorem:

Theorem 1. Suppose \mathcal{H} has level VC-dimension V . There exists a universal constant C such that uniformly for all $h \in \mathcal{H}$ the following holds with probability at least $1 - \delta$:

$$L^{iso}(h) \leq \sup_{t \in \mathbb{R}} \inf_{m \in \mathcal{M}} L_n^{thr}(m \circ h, t) + C \sqrt{\frac{V \log(n) + \log \frac{1}{\delta}}{n}}$$

The bound is a straightforward application of uniform risk bounds, provided that the VC dimension of the loss class can be found. This is a straightforward shattering argument. Suppose \mathcal{H} has level VC dimension V . Suppose there exists a set $\{(x_i, y_i)\}_{i=1}^{2V+1}$ and some t_0 such that $\{l^{thr}(h(x_i), y_i, t_0)\}_{h \in \mathcal{H}}$ can attain any labeling. Considering the two sets $\{i : y_i > t_0\}$ and $\{i : y_i \leq t_0\}$, one of these has size at least $V + 1$ by the pigeonhole principle. Call the set S . Then $\{h(x_i) - t_0 > 0\}_{h \in \mathcal{H}, i \in S}$ is shattered, which contradicts our assumption about the level VC dimension of \mathcal{H} . Hence, the VC-dimension is at most $2V$. Extending the argument to allow separate orderings in each class is similar; there are now four pigeonholes because we have to consider a separate threshold for each ordering, so $4V + 1$ points will ensure a contradiction.

3 Putting the pieces together

The ordinal regression bound is the substantial piece of the puzzle, but to make use of it have to enforce class balance. Suppose that the user provides some parameter w as a weight to favor or discourage loss in one class over another:

⁵ Level VC-Dimension is an extension of VC-dimension to real-valued functions. It is the VC dimension of $\{h - t : h \in \mathcal{H}, t \in \mathbb{R}\}$.

Definition 7. Loss balance, given a specified parameter $w > 0$, is measured by

$$l^B(\hat{y}, y; w) = \begin{cases} \max(w, 1) & \hat{y} \leq 0, y > 0 \\ -\max(1, \frac{1}{w}) & \hat{y} > 0, y \leq 0 \end{cases}$$

The VC dimension of this loss class can be analyzed in terms of the VC dimension for the underlying hypothesis space just like in the ordinal regression application, giving a similar bound. With this in hand, we are ready for the main theorem:

Theorem 2. Suppose some predefined ordering is provided. Consider the subset of \mathcal{H} that satisfies the predefined ordering up to loss d and assume that the user provides a loss balance parameter that is empirically satisfied:

$$\hat{\mathcal{H}}_d = \{h \in \mathcal{H} : L_n^{iso}(h) \leq d, L_n^B(h; w) = 0\}.$$

Assume that this set is not empty. Let \hat{h}_n be the empirical minimizer (of L_n^{01}) and \hat{h}' be the true minimizer (of L^{01}) over $\hat{\mathcal{H}}$. Then there exists a constant C such that the following holds uniformly for all $h \in \hat{\mathcal{H}}$ and for all $\phi \leq 1$ with probability at least $1 - \delta_1 - \delta_2 - \delta_3$:

$$\begin{aligned} & \max(L^{01}(h_n) - L^{01}(h'), L_n^{01}(h') - L_n^{01}(h_n)) \\ & \leq \frac{8}{n\phi} \left(4C^2V \log n + (1 + 2\phi) \log \frac{1}{\delta_1} \right) + 128\phi \left(d + C \sqrt{\frac{V \log(n) + \log \frac{1}{\delta_2}}{n}} \right) \\ & \quad + 128\phi C \max\left(w, \frac{1}{w}\right) \sqrt{\frac{V \log(n) + \log \frac{1}{\delta_3}}{n}} \end{aligned}$$

The key to interpreting this bound is to note that one tunes ϕ to optimize the value in n . While the first term on the RHS dominates, let $\phi = 1$, but when this drops below subsequent terms, ϕ can be set like $n^{-1/4}$, giving an overall convergence like $n^{-3/4}$ down to a constant times d , and then ϕ would be set like $n^{-1/2}$ thereafter. Obviously, a small d is important; that is to say, the privileged orderings should fit well to some $h \in \mathcal{H}$ under L^{iso} . However, d could be much smaller under favorable circumstances. We believe it would be possible to characterize these by extending the Mammen-Tsybakov noise conditions to L^{iso} . A proof is given in the appendix.

4 Optimizing L_n^{iso}

We now study methods to find $\inf_{h \in \mathcal{H}, m \in \mathcal{M}} L_n^{iso}(m \circ h)$, assuming a linear or RKHS hypothesis space with defined level VC-dimension. We note that for a fixed h , the optimal m can be found by means of a dynamic program. A method to construct real-valued functions with defined level VC-dimension was given by Vapnik [4, p. 359]. The technique requires the model to separate each successive example by a minimum margin. As common with zero-one loss, we relax it to hinge loss to make a convex model.

Assume that examples are sorted increasing in y_i and there are no duplicates (which require extra attention). Define

$$y_{ij} = \begin{cases} -1 & y_j \leq y_i \\ 1 & y_j > y_i \end{cases}$$

In the following formulation, C is a user-defined capacity control parameter and $\rho = 1$ can be assumed:

$$\min_{w, \xi, \zeta, l} \frac{1}{2} \|w\|^2 + C \max_i (l_i) \quad (5)$$

$$\text{s.t. for } i = 1 \dots n : \quad (6)$$

$$l_i = \sum_j \max(\rho - y_{ij}(w \cdot x_j - \xi_i), 0) \quad (7)$$

$$\text{for } i = 2 \dots n : \quad (8)$$

$$\xi_{i-1} + \rho \leq \xi_i \quad (9)$$

This is a convex quadratic programming problem, though regrettably requiring n^2 (where n is the number of examples) dummy variables to compute the max in line 7. This makes the program intractable (by machine learning standards), at least without a special solver.

The regression problem can be made more tractable by relaxing it to alternative ordinal regression formulations, such as the one proposed by Sashua & Levin [5]. Their formulation penalizes uses optimization variables to define ordered slots according to the sort order of the targets y_i , and a training example is penalized if it does not project into its slot. It can be shown that the relaxation loss is an upper bound on the unrelaxed loss. The relaxed formulation is a quadratic program with just n constraints.

5 GO-SVM

The Global-Order SVM (GO-SVM) is the name for the formulation we propose. It is simply is the usual SVM hypothesis space (thick hyperplanes) and loss (hinge), but it is simultaneously optimized with the Sashua & Levin [5] ordinal regression relaxation, with the SVM discriminant w constrained to be the same as the ordering hypothesis w . This constraint implements the restriction from \mathcal{H} (the SVM hypothesis space) to $\hat{\mathcal{H}}$ (hypotheses that satisfy an ordinal condition), as defined in Theorem 2. Loss and capacity control are traded off between the bi-objectives by means of user-selectable weights.

Capacity control in both SVM and the ordinal regression formulations is attained by the relationship between the squared norm of the predictor w and the size of the margin. In either formulation by itself, one can fix the margin size and place all capacity control in the squared norm of w , trading it off with loss. However, w serves a two-fold role in this formulation; therefore implementing different capacities for the two learning objectives requires setting the margins. Noting that the ν -SVM formulations [6] have the margin as an optimization variable, we extended the approach so that the usual tradeoff between loss and capacity is preserved.

The formulation is

$$\min_{\substack{\xi \geq 0 \\ w, b, g, \xi \\ \zeta, \rho_b, \rho_o}} \frac{1}{2} w^T w + \alpha \left(-\nu_b \rho_b + \frac{1}{n} \sum_{i=1}^n \xi_i \right) \quad (10)$$

$$+ (1 - \alpha) \left(-\nu_o \rho_o + \frac{1}{n^*} \sum_{i=1}^n |\zeta_i| \right)$$

$$\text{s.t. } \forall i, y_i(w \cdot x_i + b) \geq \rho_b - \xi_i \quad (11)$$

$$\forall i, g_{\mathcal{I}(i)} + \frac{\rho_o}{2} \leq w \cdot x_i + \zeta_i \leq g_{\mathcal{I}(i)+1} - \frac{\rho_o}{2} \quad (12)$$

Here, \mathcal{I} is an index function that returns an in-order, unique index for each distinct oracle value for in each class, and g is a vector of interval boundaries. Ordering is enforced because there are no empty intervals. The within-class ordering variant in principle requires two ordinal regressions, but in practice it can be done with a trick using the index function \mathcal{I} by creating an empty interval.

Variable w is the linear predictor, b is a bias term, ξ is the hinge loss for the classification problem, and $|\zeta|$ is hinge loss for the ordinal problem. Constant n^* is defined to control the feasible range of ν_o . It is $n^2 - n/2$ if there are not ties in the ordering.

Parameter ν_b controls the VC-dimension of the 0/1 loss class, ν_o controls the maximum VC-dimension of each subproblem in L^{thr} . Finally, parameter α is related to choosing the size of d in Theorem 2, expressed in terms of the permissiveness of the ordinal loss compared to the 0/1 loss. We do not attempt to enforce loss balance, as theory tells we should; from any computed solution, we can still apply the bound for as if we had constrained it to the value achieved by the optimum; moreover, we have never known the unconstrained optimum to have unreasonable loss balance, and we have no reason to prefer otherwise.

Like ν -SVM [6], the optimization problem can be characterized in terms of ν_b and ν_o and training data. It can be proved that the problem is primal and dual feasible for $\nu_o \in [0, 1]$, $\alpha \in [0, 1]$, and $\nu_b \in [0, 2 \min(\# \text{ positive examples}, \# \text{ negative examples})/n]$; and primal unbounded/dual infeasible otherwise. The Representer Theorem [7] holds for GO-SVM, so the solution can be expressed in terms of the dual variables and kernels can be used. We used Matlab’s interior-point convex quadratic programming solver. The baseline ν -SVM formulation was also implemented using the same solver, so that differences in numerical accuracy could not arise.

6 Evaluation

The goal of evaluation is to prove that the order oracle hypothesis space allows faster convergence than a learning formulation which considers only the labels. Since GO-SVM is an extension of standard SVM, it is a logical baseline, and we compare only to that. However, these experiments are similar to other common experiments in LUPI literature, and we will point these to the reader where appropriate. Moreover, because of the construction of the GO-SVM hypothesis spaces, it cannot outperform SVM by virtue of a richer hypothesis space. Faster convergence is the only alternative explanation. The evaluation is not intended to be a statement about the fitness of the hypothesis spaces for the learning task, but only about the ability of the learner to select the best element.

The experimental setup is to hold out a testing set and sample remaining examples for 12 random realizations of training and validation sets. The validation set is used for a set of model selection experiments, and results are reported on the test set, which is used for all experiments. Testing sets contained at least 1800 examples. The formulations have a fixed, auto-scaling parameter ν , and we use structural risk minimization to choose from a fixed set of parameters $\nu = [.1 : .1 : .9, .95]$.

The rbf kernel width (where used) is chosen from the $[.1, .25, 5]$ -quantiles of the pairwise distance of training points. The kernel parameter was chosen by a hold-out validation on the SVM experiment and re-used in the GO-SVM formulation to cut down the size of the model search. The α parameter in the GO-SVM method was chosen from $[.1, .25, .5]$.

The first evaluation is up/down prediction of the MacKey-Glass synthetic timeseries [8]. It was used in the LUPI setting (SVM+) in [1], where the authors used a 4-dimensional embedding $(x_{t-3}, x_{t-2}, x_{t-1}, x_t)$ in order to predict $x_{t+5} > x_t$. Privileged information was a 4-dimensional embedding around the target: $(x_{t+3}, x_{t+4}, x_{t+6}, x_{t+7})$. The authors compared SVM+ to SVM. We were not able to replicate their results for either SVM or SVM+, which we suspect arises from the parameters used to generate the timeseries. (We

used an integration step size of .1, with points created every 10, delay constant $\tau = 17$, and initial value .9.) We use $|x_{t+5} - x_t|$ as the order oracle. We use an RBF kernel for all experiments with this dataset.

The second evaluation is predicting binary survival at a fixed time from onset. We create synthetic datasets using the same procedure as Shiao & Cherkassky [9, personal communication], with noise level .1 and no censoring, which is given by an exponential distribution parameter $1/.01$. While censored data are an inherent aspect of survival studies, we avoid it in this case because the ordinal model can be modified to accommodate the partial information that censored examples contain; thus, it is an experiment for another day. They compare SVM, SVM+, and the Cox proportional hazards model. Privileged information for SVM+ was related to the patient’s overall survival time and whether the event time is right censored (only known to be greater than some value). We use the (absolute) difference in the fixed prediction horizon and the event time for the order oracle, and we ignore whether an example is censored. We consider only linear models.

The last evaluation is handwritten digit recognition, which was used by Vapnik and Vahist [1] for SVM+ and slightly adapted by Lapin *et al.* for their proposed LUPi method [10]. The task is to classify downsampled (10×10) MNIST images based on pixel values. Lapin added human-annotated confidence scores to training examples (available for download). We repeat the experiment using their data preparation and using their annotators’ confidence scores as the order oracle. These experiments use an RBF kernel.

6.1 Model selection

The Go-SVM formulation considers a model space of 3 dimensions: a parameter to control the complexity of the classification problem, a parameter to control the complexity of the ordinal regression problem, and a parameter that balances the loss between these two problems. All of the parameters are chosen from fixed lists, which were detailed supra. The most basic form of model selection requires choosing the best node of the grid.

We found that traditional hold-out model selection strategies are more difficult with GO-SVM. The trouble appears to be because the assumptions of structural risk minimization [4] no longer hold. In traditional SRM, the nested hypothesis spaces ensures that the loss expectation of the empirical risk minimizer (as a function of complexity) is coercive, making the selection of a minimum considerably more reliable than if it occurred at random. In our framework, there is no total ordering of hypothesis complexity. Some hypothesis spaces (defined by parameters) have good convergence, while others do not. The task is to differentiate them.

We analyzed loss surfaces with respect to various dimensions of the parameter selection grid. We used synthetic datasets where we could to generate a great number of examples. We observed that, although the surfaces were not coercive, they tended to be smooth. Since the parameters of the models have a direct interpretation, parameters that are similar should have similar performance in expectation when trained on the same set. Thus, we decided to try a Gaussian filter to smooth the validation results, and then select the minimum in a grid search. The filter was constructed apriori, and was used for all the datasets in the evaluation.

In each experiment, we computed the loss on the test set that would have been found by each of three methods:

1. A standard holdout, with the held-out validation set the same size as the training set. One might use cross-validation in practice. This is called ‘unsmooth.’
2. The Gaussian smoothing technique using the same holdout. This is called ‘smoothed.’

Table 1: Results (error rate) for all experiments. Winners are reported excluding the extended validation experiments.

experiment	ν -SVM std	ν -SVM ext	GO-SVM non-smooth	GO-SVM smoothed	GO-SVM extended
MacKey-Glass (20, 4000)	.289 (.096)	.220 (.065)	.156 (.105)	.163 (.118)	.110 (.098)
MacKey-Glass (50, 4000)	.117 (.040)	.096 (.032)	.058 (.024)	.045 (.016)	.035 (.014)
Survival (20, 1000)	.409 (.030)	.399 (.206)	.391 (.035)	.397 (.030)	.357 (.030)
Survival (40, 1000)	.322 (.032)	.311 (.029)	.300 (.027)	.287 (.024)	.275 (.026)
Survival (100, 1000)	.243 (.020)	.235 (.014)	.221 (.023)	.220 (.015)	.207 (.009)
Digits (60, 2500)	.113 (.022)	.108 (.020)	.109 (.020)	.108 (.022)	.102 (.021)
Digits (80, 2500)	.093 (.013)	.089 (.007)	.094 (.017)	.089 (.008)	.082 (.007)

3. A very large ‘oracle’ validation set which reveals how good the best hypothesis space is. This is called ‘extended.’

The Gaussian filter was of size 5x5x3, with the smallest dimension corresponding to the α parameter. (In experiments using a kernel parameter, we used one found via the SVM model search, so this was not a grid search parameter.) It has the property that any projection along coordinate directions of the filter is Gaussian. This was convolved with the tensor of validation scores using zero-degree smooth extrapolation; that is, the tensor is padded out with constants which are the same as the nearest true element of the tensor. The convolution gives a new tensor that is the same dimension as the un-smoothed tensor.

We also considered two alternative validation scenarios: first, selecting a model based on the un-smoothed tensor, and second, investigating the effect of having a much larger validation set available. The large validation set is intended to point out the gap between the best hypothesis spaces that can be created using the ordinal constraint technique, and the one which can in practice be selected. We point out that many LUPI research papers require validation sets that would not ordinarily be a reasonable split between training and testing data (for example [11, 1]). Each row of the table gives the size of the training and test sets. The columns give the model selection procedure.

6.2 Conclusions

A table of results is given at Table (6.1). As a reminder, sizes for training and testing are given in parenthesis with the experiment name. The std, non-smooth, and smoothed methods used a validation set the same size as the training set. We note first that the gap between the extended validation model selection and the performance of the typical technique is larger for GO-SVM than for standard SVM. This is a blessing in that we have the opportunity to find a better model, but also a curse in that the variance is higher. It appears that this strain of LUPI methods is bound by model selection issues. The Gaussian smoothing approach seems to have been effective on the Digits dataset, and certainly did not hinder performance significantly where the un-smoothed model selection turned out to be superior.

The MacKey-Glass dataset is the only one which has strongly significant results. Although the gains in other datasets are small, the fact that they are supported by theory implies they should not be overlooked. Moreover, they are consistent with results reported by other authors. There comparisons with other works for the MacKey-Glass and Digits experiments. The MacKey-Glass experiment appeared in the original SVM+ paper [1]. They report, based on a training set of 100 examples, that SVM had an error rate of .052, whereas

the best SVM+ formulation was at .048. Furthermore, this was based on a validation set of size 500. We have reached that level of performance with considerably less data. The Digits experiment is intended to replicate one in [11]. We specifically replicated the experiment in which conditional probability weights were created by humans with an intent to help a machine. This task is well-suited to the order invariance that GO-SVM is built on, as humans have a fundamentally ordinal notion of confidence. In that study at a sample size of 80, the difference between the best and worst methods under study was about .01—.073 compared to .083 (approximately). The size of the validation set in use, would be comparable to our extended experiment. Their best method, however, did not use human weights. In their experiment, the human weights information improved over SVM by about .003, whereas our gain is .006.

In conclusion, the fact that the formulation can find faster-converging models than formulations which don't consider order information supports the underlying theory. It appears likely the order information is helpful in scenarios when the prediction task discretizes some continuous attribute, such as in the timeseries and survival prediction tasks.

7 Previous work

The original SVM+ paper [1] touched off a fair amount of research in the area. Most research, with limited exceptions, has focused on developing and evaluating formulations [12, 13, 14, 15, 15] rather than attempting to develop theory to understand when and why such a technique might be useful.

Pechyony et. al. [16] analyze the SVM+ algorithm in terms of variance bounds. While it shares with this work a major emphasis on variance bounds, that work considers the SVM+ loss function as given and derives bounds for it, whereas this paper works the other way in attempting to derive a formulation based on the bound.

Lapin et. al. [11] propose weighting examples based on class conditional probability, and is most directly similar to the ideas proposed here. Intuitively, the method encourages a learner to prioritize performance on the easy examples over the hard examples. Unfortunately, the theoretical motivation for departing from empirical risk minimization takes a tenuous path through SVM+ [1, 16], namely that SVM+ is reducible to weighted learning. The heart of their method is based on a loss function based on weights interpreted as conditional probability; however, a theoretical analysis is not provided. Our is somewhat more general in allowing order in variances.

References

- [1] Vladimir Vapnik and Akshay Vashist. 2009 special issue: A new learning paradigm: Learning using privileged information. *Neural Netw.*, 22:544–557, July 2009.
- [2] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 12 1999.
- [3] Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 02 2004.
- [4] V.N. Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, 1998.
- [5] Amnon Shashua and Anat Levin. Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 937–944, 2002.

- [6] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Comput.*, 12(5):1207–1245, May 2000.
- [7] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, volume 2111 of *Lecture Notes in Computer Science*, pages 416–426. Springer Berlin Heidelberg, 2001.
- [8] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 511–520, Sep 1997.
- [9] Han-Tai Shiao and Vladimir Cherkassky. Learning using privileged information (LUPI) for modeling survival data. In *2014 International Joint Conference on Neural Networks, IJCNN 2014, Beijing, China, July 6-11, 2014*, pages 1042–1049, 2014.
- [10] Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: SVM+ and weighted SVM. *Neural Networks*, 53:95–108, 2014.
- [11] Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: SVM+ and weighted SVM. *Neural Networks*, 53:95–108, 2014.
- [12] Jixu Chen, Xiaoming Liu, and Siwei Lyu. Boosting with side information. In *Computer Vision - ACCV 2012 - 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I*, pages 563–577, 2012.
- [13] Ziheng Wang, Tian Gao, and Qiang Ji. Learning with hidden information using a max-margin latent variable model. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1389–1394, Aug 2014.
- [14] Shereen Fouad, Peter Tino, Somak Raychaudhury, and Petra Schneider. Learning using privileged information in prototype based models. In Alessandro E.P. Villa, Włodzisław Duch, Péter Érdi, Francesco Masulli, and Günther Palm, editors, *Artificial Neural Networks and Machine Learning – ICANN 2012*, volume 7553 of *Lecture Notes in Computer Science*, pages 322–329. Springer Berlin Heidelberg, 2012.
- [15] Ziheng Wang and Qiang Ji. Classifier learning with hidden information. *Computer Vision and Pattern Recognition, IEEE Conference on*, 2015.
- [16] D. Pechyony and V. Vapnik. On the theory of learning with privileged information. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, 2010.
- [17] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification : a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.
- [18] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

8 Appendix

The first result is a bridge between the variance conditions and the metric balls of hypotheses we can actually define. This result shows that the uniform variance conditions can be relaxed by a small constant, depicted here as $d(n)$, at the expense of the rate of convergence when the bound is small.

Theorem 3. Suppose there is a loss class $\mathcal{F} = \{l \circ h : h \in \mathcal{H}\}$ with VC-dimension V , and let f' attain $\inf_{f \in \mathcal{F}} \mathbb{E}[f]$. There is a uniform constant C such that if the following holds uniformly for all $f \in \mathcal{F}$,

$$\text{Var}[f - f'] \leq \frac{1}{h} \mathbb{E}[f - f'] + d(n)$$

then for any $\phi \leq h \leq 1$, the following holds uniformly with probability $1 - \delta$:

$$\begin{aligned} & \max \left(\mathbb{E}[f_n - f'], \mathbb{E}[f' - f_n] \right) \\ & \leq \frac{8}{n\phi} \left(4C^2 V \log n + (1 + 2\phi) \log \frac{1}{\delta} \right) + 32\phi d(n). \end{aligned}$$

Proof. I follow the definitions and notation in Bucheron *et al.* [17, Theorem 5.5]. In their notation it is straightforward to show that $w(r) \leq \sqrt{\frac{r}{\phi} + d}$. For VC-classes, it can be proved that $\psi(x) \leq Cx\sqrt{\frac{V}{n} \log n}$ [18]. The risk bound depends on the solution of a fixed-point equation. Let ϵ^* be the solution of $r = \psi(w(r))$. Let $\epsilon' = C^2 \frac{V \log n}{n\phi} + \phi d$. The following analysis shows that $\epsilon' \geq \psi(w(\epsilon'))$, which implies $\epsilon^* \leq \epsilon'$.

$$\psi(w(\epsilon')) = C \left(\frac{C^2 V \log n}{\phi^2 n} + 2d \right)^{\frac{1}{2}} \left(\frac{V \log n}{n} \right)^{\frac{1}{2}} \quad (13)$$

$$\leq C \left(\left(\frac{C^2 V \log n}{\phi^2 n} \right)^{\frac{1}{2}} + \frac{1}{2} \left(\frac{C^2 V \log n}{\phi^2 n} \right)^{-\frac{1}{2}} 2d \right) \left(\frac{V \log n}{n} \right)^{\frac{1}{2}} \quad (14)$$

$$= C^2 \frac{V \log n}{n\phi} + \phi d = \epsilon'. \quad (15)$$

At step 14 I used that the first-order approximation of the square root is an upper bound. The bound ϵ' can be substituted wholesale into the bound given by Bucheron in the statement of the theorem, which gives the theorem. \square

The next step is to show that the combination of the ordinal constraint and the balance constraint are sufficient to bound the variance diameter of the subset of a hypothesis space that satisfies those conditions.

Lemma 2. Suppose that an ordering and loss balance parameter w are provided and that $f, g \in \hat{\mathcal{H}}$. That is, $L^{iso}(f) \leq D_0$, $L^{iso}(g) \leq D_0$, and $L^B(f) \leq B_0$ and $L^B(g) \leq B_0$. Finally, Suppose $L^{01}(g) \leq L^{01}(f)$. Then $\mathbb{E}[|l^{01}(f(X), Y) - l^{01}(g(X), Y)|] \leq \mathbb{E}[l^{01}(f(X), Y) - l^{01}(g(X), Y)] + 4(b + d)$.

Proof. For the purposes of the proof, we decompose the expectation by class. Let $P_P = P(X|Y = 1)$ and $P_N = P(X|Y = -1)$. Let $p_P = P(Y = 1)$ and $p_N = P(Y = -1)$. Similarly, let L_P and L_N be loss functions defined by conditional expectations. The outline of the argument is shown below. We will expand each line subsequently.

$$\mathbb{E}_P |l^{01}(f(X), Y) - l^{01}(g(X), Y)| \quad (16)$$

$$= \mathbb{E}_{P_P} |l^{01}(f(X), 1) - l^{01}(g(X), 1)| p_P + \mathbb{E}_{P_N} |l^{01}(f(X), -1) - l^{01}(g(X), -1)| p_N \quad (17)$$

$$\leq \mathbb{E}_{P_P} |l^{01}(f(X), 1) - l^{01}(g(X), 1)| p_P + \mathbb{E}_{P_N} |l^{01}(f(X), -1) - l^{01}(g(X), -1)| p_N + 4D_0 \quad (18)$$

$$\leq \mathbb{E}_{P_P} [l^{01}(f(X), 1) - l^{01}(g(X), 1)] p_P + \mathbb{E}_{P_N} [l^{01}(f(X), -1) - l^{01}(g(X), -1)] p_N + 4(D_0 + B_0) \quad (19)$$

$$= L^{01}(f) - L^{01}(g) + 4(D_0 + B_0) \quad (20)$$

We begin by proving the inequality at line 18. Consider only the positive class for a moment. Let δ_f be the decision (or margin, as a simple extension) boundary for f and δ_g the boundary for g . Then l^{01} is 1 for $f(X) \leq \delta_f$ and 0 otherwise.

We have noted the triangle inequality relationship between D and L^{iso} . Suppose $L^{iso}(f) \leq D_0$ and $L^{iso}(g) \leq D_0$, then $D(f, g) \leq 2D_0$. Let m_f and m_g be monotone functions as defined (implicitly) in L^{iso} . Then $m_g^f = m_g^{-1}m_f$ is a continuous monotone function which makes the metric relationship true. We will first show

$$\mathbb{E}_{P_P} |\mathbb{1}_{f(X) \leq \delta_f} - \mathbb{1}_{g(X) \leq \delta_g}| \leq |\mathbb{E}_{P_P} [\mathbb{1}_{f(X) \leq \delta_f} - \mathbb{1}_{g(X) \leq \delta_g}]| + 2D_0 \quad (21)$$

$$\Leftrightarrow \mathbb{E}_{P_P} [\mathbb{1}_{f(X) \leq \delta_f \wedge g(X) > \delta_g}] + \mathbb{E}_{P_P} [\mathbb{1}_{f(X) > \delta_f \wedge g(X) \leq \delta_g}] \quad (22)$$

$$\leq |\mathbb{E}_{P_P} [\mathbb{1}_{f(X) \leq \delta_f \wedge g(X) > \delta_g}] - \mathbb{E}_{P_P} [\mathbb{1}_{f(X) > \delta_f \wedge g(X) \leq \delta_g}]| + 2D_0 \quad (23)$$

Rewriting concisely, we wish to show $a + b \leq |a - b| + 2D_0$. In fact, this holds because either $a \leq D_0$ or $b \leq D_0$, which can be proved in the following way: Expanding a we have:

$$\mathbb{E}[\mathbb{1}_{f(X) \leq \delta_f \wedge g(X) > \delta_g}] = \mathbb{E}[\mathbb{1}_{f \leq \delta_f \wedge g > \delta_g \wedge g > m_g^f(\delta_f)}] + \mathbb{E}[\mathbb{1}_{f \leq \delta_f \wedge g > \delta_g \wedge g \leq m_g^f(\delta_f)}] \quad (24)$$

$$\leq D_0 + \mathbb{E}[\mathbb{1}_{g > \delta_g \wedge g \leq m_g^f(\delta_f)}] \quad (25)$$

The expectation term in line 25 is 0 if $m_g^f(\delta_f) \leq \delta_g$. Repeating the procedure for b shows that the corresponding term is 0 if $m_g^f(\delta_f) \geq \delta_g$. Since one of those conditions must be true, at least one of a or b is bounded by d , which proves line 21. A bound for the negative class is identical. This proves inequality 18.

To prove inequality 19, the assumptions on class balance are needed. Since we assumed that $L(f) > L(g)$, then either (or both) $L_P(f) > L_P(g)$ or $L_N(f) > L_N(g)$. If both are true, then the desired inequality (19) is trivial. Suppose that $L_P(f) > L_P(g)$ and $L_N(f) \leq L_N(g)$.

$$|L_P(f) - L_P(g)|p_P + |L_N(f) - L_N(g)|p_N \quad (26)$$

$$= |a - b| + |c - d| \quad (27)$$

$$= a - b + d - c \quad (28)$$

$$= a - b + c - d + 2(d - c) \quad (29)$$

$$\leq a - b + c - d + 2w(b - a) + 4B_0 \quad (30)$$

$$\leq (L_P(f) - L_P(g))p_P + (L_N(f) - L_N(g))p_N + 4B_0 \quad (31)$$

where we used that $b - a < 0$, $|wa - c| \leq B_0$, $|wb - d| \leq B_0$, and $w > 0$. The proof if $L_N(f) > L_N(g)$ and $L_P(f) \leq L_P(g)$ uses that $|a - \frac{1}{w}c| \leq B_0$ and $|b - \frac{1}{w}d| \leq B_0$. \square